行政院國家科學委員會專題研究計畫 成果報告

2×2 列聯表之穩健診斷
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 99-2118-M-004-002-
執 行 期 間 ：99 年 08 月 01 日至 100 年 09 月 30 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人 ：鄭宗記

計畫參與人員：碩士級-專任助理人員：許斯淵

報 告 附 件 ：出席國際會議研究心得報告及發表論文

公 開 資 訊 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 100 年 11 月 29 日

中 文 摘 要 ： 吾人在本計畫中提出對邏輯斯迴歸模型的最小削減偏差 (minimum trimmed deviances ，MTD)穩健估計方法。另外針對共同勝算迴歸模型又提出最小削減最大成對偏差 (minimum trimmed maximum-dual deviances ，MTMdD)估計方法，此為對 MTD 的一個延伸，藉此可以偵測數個 2×2 列聯表分析時其可能的離群值。

中文關鍵詞： 2×2 列聯表、邏輯斯迴歸模型、最小削減偏差估計

英 文 摘 要 ：

英文關鍵詞：

Robust diagnostics for analyzing several 2×2 contingency tables

Tsung-Chi Cheng*

Abstract

We propose a minimum trimmed deviances (MTD) estimator for the robust estimation of the logistic regression model. An adoption of MTD, called the minimum trimmed maximum-dual deviances (MTMdD) estimator, is applied to estimate the common odds regression model and hence to identify outlying tables when some possible outliers exist among several 2×2 contingency tables. We will illustrate the methods by analyzing a real data example in the literature.

KEY WORDS: 2×2 contingency tables; logistic regression model; minimum trimmed deviances estimator; Robust diagnostics

1 Introduction

The logistic regression model is one of the most powerful tools in data analysis for medical and epidemiologic studies. The maximum likelihood estimation (MLE) is the popular approach to calculate the coefficient estimation for a logistic regression model, but it is sensitive to outlying responses and extreme points in the design matrix. Outliers in the design matrix are called leverage points or influential points in the regression literature. Both types of outliers can spoil the maximum likelihood fit for a logistic regression analysis. However, such outlying observations are hard to identify, because they do not always show up in the usual residual plots.

In the context of categorical variables for the contingency tables, outliers have not been studied frequently. Outlier identification and accommodation in contingency tables have been discussed by Kotze and Hawkins (1984), Simonoff (1988), and Yick

*Department of Statistics, National Chengchi University, 64 ZhihNan Road, Section 2, Taipei 11605, Taiwan. E-mail: chengt@nccu.edu.tw

and Lee (1998), with these works focusing on the identification of outliers in a single contingency table. In practice, there exist multiple 2×2 contingency tables to be analyzed. One of the principal advantages of using the logistic regression model is that it encourages a quantitative description of how changes in risk associated with one factor are modified by the interaction effects of other risk or nuisance variables. A generalization of the Mantel-Haenszel estimator to non-constant odds ratios was proposed by Davis (1985) as an alternative to the conditional maximum likelihood for fitting log odds ratio regression models to sets of sparse 2×2 tables such as those that arise in case-control studies (Breslow, 1976; Breslow and Cologne, 1986). The presence of interaction effects in a series of 2×2 contingency tables depends systematically on the variables used for strata formation. Hence, a model is considered for this purpose, in which the log relative risk is assumed to change linearly over a strata.

In this project we first propose a minimum trimmed deviances (MTD) estimator for the robust estimation of the logistic regression model, which is inspired by the least trimmed squares (LTS) technique. An adoption of MTD, called the minimum trimmed maximum-dual deviances (MTMdD) estimator, is then used to estimate the common odds regression model and hence to identify outlying tables for the analysis of several 2×2 contingency tables. The fast algorithm is adapted to find both resulting estimators. The proposed procedure is illustrated by using real data analysis.

2 The logistic regression model

Let there be n binomial observations of the form y_i/m_i, $i = 1, 2, \ldots, n$, where $E(y_i) = m_i\pi_i$, and π_i is the success probability corresponding to the ith observation. The binomial distribution for a fixed number of trials is determined by the probability π of success. Both the mean and the variance depend only on π_i and the known number m_i of trials.

For each y_i we know the number of trials m_i, and in addition there is an associated vector of $p + 1$ predictors $\boldsymbol{x_i}$. Assuming that the probability of success depends on

$\boldsymbol{x_i}$, then the probability function of y_i can be written as:

$$\begin{aligned} \pi_i &= P(y_i = 1) = f(\boldsymbol{x}; \boldsymbol{\beta}) \\ &= \frac{\exp(\boldsymbol{x_i^T \beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is a $(p+1) \times 1$ unknown parameters vector. Note that $0 \leq \pi_i \leq 1$ for all values of $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$. The log odds ratio is:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

which is linear in the parameters $\boldsymbol{\beta}$. The logistic regression may be viewed as a non-linear model with heterosceddastic errors - that is:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim bin(m_i, \pi_i),$$

where $E(\epsilon_i) = m_i \pi_i$ and $\text{Var}(\epsilon_i) = m_i \pi_i (1 - \pi_i)$. These parameters are readily estimated using the method of MLE (see McCullagh and Nelder (1989) for details), which is given by maximizing

$$L(\boldsymbol{\eta}; \boldsymbol{y}) = \sum_{i=1}^{n} l(\pi(\boldsymbol{x}_i^T \boldsymbol{\beta}); y_i), \tag{2}$$

where $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $l(\pi(\boldsymbol{x}_i^T \boldsymbol{\beta}); y_i)$ denotes the log likelihood for the ith case.

Once the estimator of $\boldsymbol{\beta}$ is obtained, denoted by $\hat{\boldsymbol{\beta}}$, the estimated value of the model is:

$$\hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}. \tag{3}$$

The fitted probabilities $\hat{\pi}_i$ can now be found using $\hat{\pi}_i = \exp(\hat{\eta}_i)/[1 + \exp(\hat{\eta}_i)]$. In the logistic regression there are several possible ways to measure the difference between the observed and the fitted values. One of them is the signed square-root deviance residual, which is defined as:

$$r_i = sign(y_i - \hat{y}_i) \sqrt{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(m_i - y_i) \log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right)}, \tag{4}$$

where "*sign*" denotes the sign of $(y_i - \hat{y}_i)$. When the response is binary, the signed square-root deviance residual becomes:

$$r_i = sign(y_i - \hat{\pi}_i) \sqrt{-2[y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)]}. \tag{5}$$

The deviance residual provides information about how well the model fits each particular observation (see Hilbe, 2009).

3 Minimum trimmed deviances estimator

In this section we propose the minimum trimmed deviances (MTD) estimator for the robust estimation of the logistic regression model, which is originally adapted from the LTS estimator. Instead of summing up all the squared residuals in the ordinary least squares estimation for a linear regression model, the LTS only considers to include the first q of the smallest squared residuals in the summation.

Pregibon (1982) shows how the MLE is related to the minimum deviance estimation (MDE) for a logistic regression model. As the log likelihood function is defined up to an additive constant by $l(\eta, y)$, the deviance function is defined as:

$$d(\pi_i; y_i) = -2\{l(\eta_i; y_i) - l_{\max}(\eta_i; y_i)\},$$

where $\pi_i = \pi(\eta_i) = \exp(\eta_i)/(1 + \exp(\eta_i))$, and $l_{\max}(\eta_i; y_i)$ is the maximum of $l(\eta_i; y_i)$ with respect to η_i. As $l_{\max}(\eta_i; y_i)$ is maximized over η_i and is a function of y_i alone, it is constant with respect to η_i. Thus, maximizing $L(\boldsymbol{\eta}; \boldsymbol{y})$ of (2) is formally equivalent to minimizing $D(\boldsymbol{\pi}; \boldsymbol{y})$, and so the MLE, $\hat{\boldsymbol{\beta}}$, satisfies the minimization of

$$\sum_{i=1}^{n} d(\pi(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}); y_i). \tag{6}$$

Let $\boldsymbol{\beta}_q$ denote the parameters for a specific value of q. If $\mathcal{Q}$ denotes the subset with q cases and the corresponding data are denoted by $\boldsymbol{y}_q$ and $\boldsymbol{X}_q$, then let

$$d_{(1),\mathcal{Q}} \leq d_{(2),\mathcal{Q}} \leq \cdots \leq d_{(n),\mathcal{Q}}, \tag{7}$$

where $d_{(i)}$ denotes the ith-ordered deviance residual and $d_{i,\mathcal{Q}} = d(\pi(\boldsymbol{x}_i^T \boldsymbol{\beta}_q); y_i)$. Instead of adding all the deviance residuals as in (6), we denote the MTD estimator evaluated at q as:

$$\min_{\beta_q} \sum_{i \in \mathcal{Q}} d_{(i)}(\pi(\boldsymbol{x}_i^T \boldsymbol{\beta}_q); y_i). \tag{8}$$

The resulting MTD estimator evaluated at q is denoted by $\hat{\boldsymbol{\beta}}_q$, which corresponds to the MTD estimator based on the subset $\mathcal{Q}$. The corresponding signed square-root deviance residuals (4) and (5) replace the estimated model (3) by using $\hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_q$ for $i = 1, 2, \ldots, n$.

3.1 The fast algorithm

We rely on the fast algorithm of Rousseeuw and van Driessen (1999) to obtain the MTD estimate of $\boldsymbol{\beta}_q$. The fast algorithm obtains a small subset of the data, and then the observations of the subset are augmented and updated in such a way that outliers are unlikely to be included. The basic idea behind the FAST algorithm consists of carrying out many two-step procedures: a trial step followed by a refinement step (the so-called Concentration step).

To carry out the fast algorithm for MTD, in the trial step a subsample of size s is selected randomly from the data and then the model is fitted to that subsample in order to get a trial ML estimate in this paper. Here, the subsample size, s, can be any values between p and q - for example, the so-called elemental sets are used in Rousseeuw and Van Driessen's papers. The larger the size is of the initial subset, the higher the probability is of the subset including outliers. Since categorical covariates often exist in practice and/or the binary response discussed in the subsequent subsection, a small initial subset may lead to a singular design matrix or a perfect fit. When either or both situations occur, we use a larger initial subset to avoid computational failure.

The refinement step is based on the so-called concentration procedure: (a) the cases with the q smallest deviance residuals based on the current estimate are found, starting with the trial MLE as initial estimator; (b) fitting the model to these q cases yields an improved fit. Repeating (a) and (b) leads to an iterative procedure. The convergence is always guaranteed after a finite number of steps since there are only a finite many q-subsets out of $\binom{n}{q}$ (Müller and Neykov, 2003). The one with the smallest value of (8) is then an approximate to the solution of MTD.

4 Robust diagnostics for analyzing several 2×2 contingency tables

The odds ratio is an important feature extensively used by practitioners in different applied areas. Breslow (1976) proposes a model for stratified case-control data that allows a variation of the odds ratio with variables used for stratification. He suggests using the conditional maximum likelihood estimator to estimate the parameters of the model. Davis (1985) presents an estimator that is like the conditional maximum likelihood estimation. It is a generalization of the estimator proposed by Mantel and Haenszel (1959) for the common odds ratio in a series of 2×2 tables whose asymptotic and small-sample properties have been studied extensively. However, both approaches based on the maximum likelihood are known to be sensitive to outliers.

Consider a series of K independent 2×2 contingency tables, with the data in the kth table denoted as shown in Table 1. Suppose that each y_{jk} follows a binomial distribution with parameters n_{jk} and π_{jk} ($j = 1, 2$; $k = 1, 2, \ldots, K$). Let n_{jk} denote the total number of observations in the ith treatment and kth center, and let π_{jk} denote the success probability at treatment level x_{jk} in the kth center, where x_{jk} is the treatment indicator with $x_{jk} = 1$ representing treatment 1 and $x_{jk} = 0$ is for treatment 2. Such data arise, for example, from a stratified prospective or retrospective study of the relationship between a single disease and a single dichotomous risk factor.

Note that for each k, the row sums, n_{1k} and n_{2k}, are fixed by design. Let $\pi_k(x_{jk})$ denote the probability that a patient at center k responds to treatment x_{jk}. The data from a case-control study with a single dichotomous risk factor are often stratified into 2×2 tables using some variables x. In this paper we focus on the following stratified logit model formulation of the dependence of responses on treatment in order to assess the homogeneity of odds ratios:

$$\log\left(\frac{\pi_k(x_{jk})}{1-\pi_k(x_{jk})}\right) = \alpha_k + \beta_k x_{jk}, \tag{9}$$

where α_k reflects the stratum effect, and $\theta_k = \exp(\beta_k)$ is the kth stratum treatment-response odds ratio. The estimation for model (9) has been discussed in Breslow (1976) and Davis (1985). The common odds ratio model, which assumes $\beta_k = \beta$ for

all k, is widely applied to such data as Table 1 (Gart 1971; Breslow and Day 1980; Hirji *et al.* (1996); Bagheri *et al.* (2011)).

4.1 Minimim trimmed maximum-dual deviances estimator

When there are possible outlying tables in the data, we extend the MTD estimator to obtain the robust estimation for a common odds regression model.

Each strata k or kth 2×2 table results in two cases when fitting model (9). One is for $x_{jk} = 1$, and the other for $x_{jk} = 0$. Their corresponding deviance residuals are denoted by $d_{1k} = d(x_{jk} = 1)$ and $d_{0k} = d(x_{jk} = 0)$, respectively. Both d_{0k} and d_{1k} are defined as the square of (4) but with the MTD estimator. It is then unnatural to trim any one case for the data of this kind when applying the MTD estimator to model (9). The exclusion of any 2×2 table requires excluding two observations at the same time. Let $\mathcal{Q}$ denote the subset with q 2×2 contingency tables. We first consider the maximum of each pair of deviance residuals, denoted by $d^*_{k,\mathcal{Q}} = \max(d_{0k,\mathcal{Q}}, d_{1k,\mathcal{Q}})$, for the kth 2×2 table, and then order the following deviance residuals:

$$d^*_{(1),\mathcal{Q}} \leq d^*_{(2),\mathcal{Q}} \leq \cdots \leq d^*_{(K),\mathcal{Q}}, \tag{10}$$

which decide the order of each table to contribute to the MTD criterion.

The minimum trimmed maximum-dual deviances (MTMdD) estimator for model (9) with common odds ratios, β, is defined by:

$$\min_{\beta} \sum_{k=1}^{q} d^*_{(k),\mathcal{Q}}. \tag{11}$$

where $d^*_{(k)}$ denotes the ith-ordered deviance residual of (10). Here, $[k/2]+1 \leq q \leq K$. The resulting MTMdD estimator evaluated at q is also denoted by $\hat{\beta}_q$ for the brevity of notations. This corresponds to the MTD estimator corresponding to those q tables.

The fast algorithm of section 3.1 is then applied to find the MTMdD estimator, but the re-sampling scheme is based on K tables rather than $2 \times K$ observations.

5 Real data illustration: Oxford childhood cancer survey data

This section uses some real data examples in the literature to illustrate the performance of the proposed approach. Kneale (1971) and Breslow and Day (1980) consider data from retrospective studies of the relationship between obstetric radiation and childhood cancer in the Oxford Childhood Cancer Survey. Cases and controls (corresponding to either dying or not dying of childhood cancer, respectively) are each classified according to whether the mother had been X-rayed during gestation. The covariates are year of birth and age at death. Tsujitani and Koch (1991) apply part of this data set to show the residual plots for the log odds ratio regression models. Davis (1985) compares the Mantel-Haenszel generalization with a conditional or unconditional maximum likelihood for these data in utero radiation and childhood cancer incidence, with stratification into 120 categories of age $\times$ year of birth (Breslow and Day, 1980, Appendix II). Breslow and Day (1980, Chapter 6) discuss several kinds of logistic regression to fit these data.

We herein use the zero degree model shown in Table 6.17 of Breslow and Day (1980, p. 242), which compares the log relative risk and its interaction with year of birth, depending on the degree of polynomial adjustment for age and year. Table 2 shows the estimation results using MLE and MTMdD, in which both yield slightly different conclusions in terms of the values of the estimates and their significance.

Figure 1 shows the deviance residual plots resulting from both approaches, in which the solid and dashed lines indicate the mother had been X-rayed or not, respectively, for each table. We observe that the solid lines for most tables represent the maximum of the absolute signed square-root deviance residuals, r_k^*, for the kth table. This shows whether the mother having been X-rayed during gestation can have different impacts on the cases or controls. Hence, it also explains that obstetric radiation is an important factor on childhood cancer. The MTMdD estimator identifies the last one as an outlying table for these data, where all observations appear inlying by using MLE. The patterns for both plots are relatively consistent.

References

Bagheri, Z., Ayatollahi, SMT, Jafari, P. Comparison of three tests of homogeneity of odds ratios in multicenter trials with unequal sample sizes within and among centers, BMC Medical Research Methodology 2011; **11**:58.

Bianco AM, Martínez EJ. Robust testing in the logistic regression model. *Computational Statistics and Data Analysis* 2009; **53**:4095-4105.

Breslow, N. Regression analysis of the log odds ratio: A method for retrospective studies. Biometrics 1976; **32**:409-416

Breslow NE, Cologne J. Methods of estimation in log odds ratio regression models. *Biometrics* 1986; **42**:949-954.

Breslow NE, Day, NE. *Statistical Methods in Cancer Research I: The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer, 1980.

Davis LJ. Generalization of the Mantel-Haenszel estimator to nonconstant odds ratios. *Biometrics* 1985; **41**:487-495.

Gart JJ. Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals. *Biometrika* 1970; **57**:471-475.

Hilbe JM. *Logistic Regression Models*. Boca Raton, FL: Chapman & Hall/CRC Press, 2009.

Hirji KF, Vollset SE, Reis IM, Afifi AA. Exact tests for interaction in several 2×2 tables. *Journal of Computational and Graphical Statistics* 1996; **5**:209-224.

Kneale GW. Problems arising in estimating from retrospective survey data the latent period of juvenile cancers initiated by obstetric radiography. *Biometrics* 1971; **27**:563-90.

Kotze TJ, and Hawkins DM. The identification of outliers in two-way contingency tables using 2×2 subtables. *Applied Statistics* 1984, **33**:215-223.

McCullagh P, Nelder JA. *Generalized Linear Models*. London: Chapman and Hall, 1989.

Müller CH, Neykov NM. Breakdown points of the trimmed likelihood and related estimators in generalized linear models. *Journal of Statistical and Planning Inference* 2003; **116**:503-519.

Pregibon D. Resistant fits for some commonly used logistic models with medical applications. *Biometrics* 1982; **38**:485-498.

Rousseeuw PJ, van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999; **41**:212-223.

Simonoff JS. Detecting outlying cells in two-way contingency tables via backwards-stepping. *Technometrics* 1988; **30**:339-345.

Tsujitani M, Koch GG. Residual plots for log odds ratio regression models. *Biometrics* 1991; **47**:1135-1141.

Yick JS, Lee AH. Unmasking outliers in two-way contingency tables. *Computational Statistics and Data Analysis* 1998; **29**:69-79.

Table 1. Summary of data from the kth 2×2 contingency tables

	Response		
	Success ($Y = 1$)	Failure ($Y = 0$)	Total
Treatment 1 (x_{1k})	y_{1k}	$n_{1k} - y_{1k}$	n_{1k}
Treatment 2 (x_{2k})	y_{2k}	$n_{2k} - y_{2k}$	n_{2k}
Total	t_k	$t_k - n_k$	n_k

Table 2. Estimation results for the Oxford childhood cancer survey data.

	MLE			MTMdD		
	Est.	Std Err	p-value	Est.	Std Err	p-value
Intercept	-0.064	0.020	0.001	-0.049	0.023	0.029
Xrayed	0.511	0.056	<0.001	0.407	0.065	<0.001
Xrayed:Year	-0.034	0.014	0.011	-0.060	0.016	<0.001
Res dev	119.35			40.32		
df	237			177		

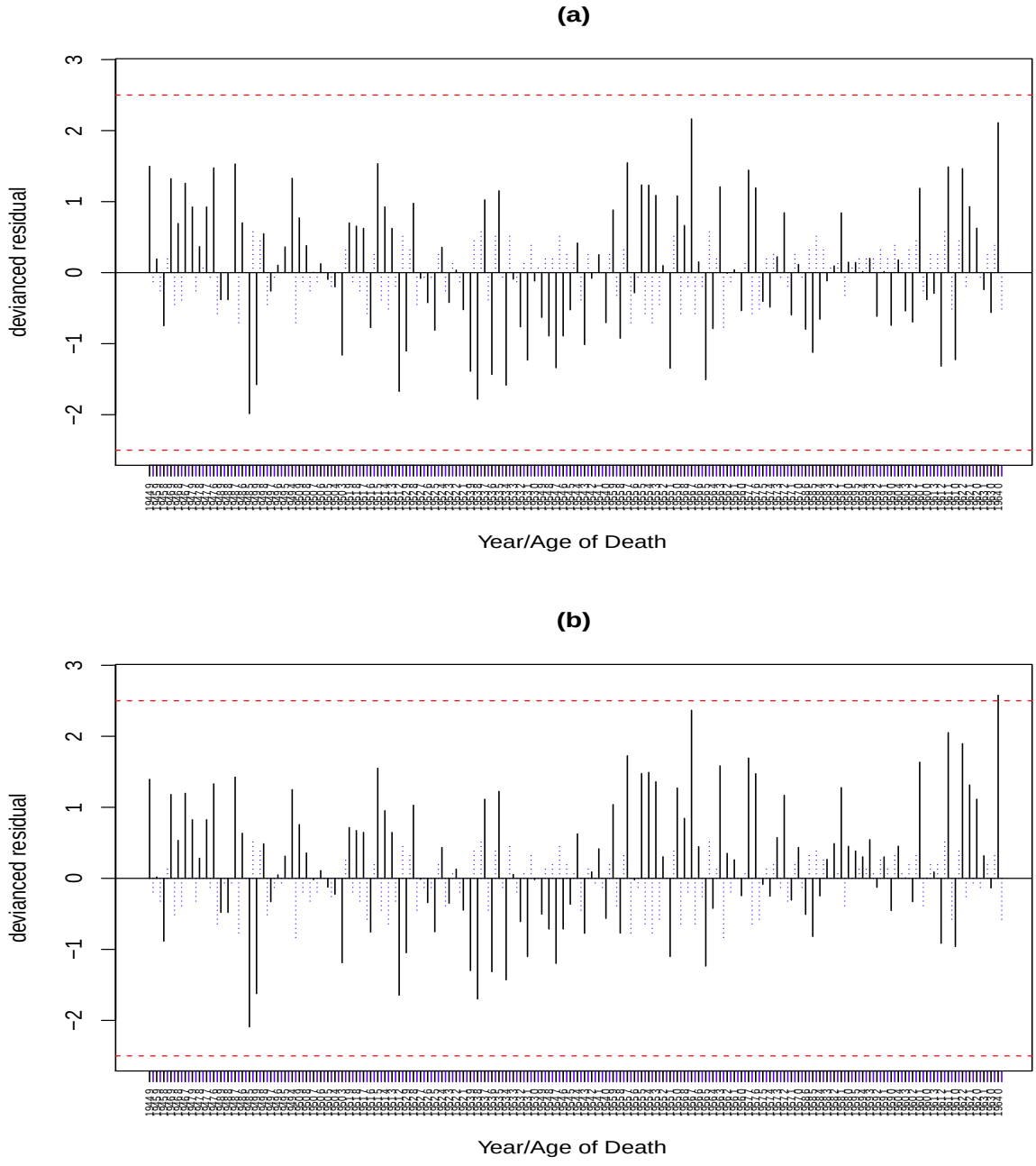

Figure 1: Deviance residual plots for the Oxford childhood cancer survey data: (a) (a) MLE; (b) MTMdD.

出席國際學術會議心得報告

計畫編號	NSC99-2118-M-004-004
計畫名稱	２×２列聯表之穩健診斷
出國人員姓名 服務機關及職稱	鄭宗記　政治大學統計系 副教授
會議時間地點	2011/09/14-17，泰國曼谷
會議名稱	2011 IEEE International Conference on Quality and Reliability
發表論文題目	On the Hotelling T^2 control chart for the vector autoregressive process

一、參加會議經過

本人於 9 月 14 日出發至泰國曼谷，參與本年度之 T IEEE International Conference on Quality and Reliability，於 9 月 16 日下午發表上述之論文；並於 9 月 17 日上午擔任一個場次的主持人。與會期間並出席其他場次論文發表，收獲甚多。

會議期間與其他國家學者多有交誼，其中與會議主辦者來自香港城市大學的謝旻教授幾次討論，得以知道此會議之特色，與舉辦會議之辛苦。

二、與會心得

本次大會參加者因地緣關係，甚多來自亞洲國家；且其因會議性質與主題，來自產官學界之身分者皆有。就各方面而言，此次的與會有相當大的收穫。在個人研究方面，此行與幾位學者的接觸，與聆聽演講，都引發一些未來可能的研究主題與方向。

由前述與謝教授之交談得知，此會議每兩年由亞洲國家輪流舉行，以國內在此一相關主題的研究成果及能量，若能爭取此會議之主辦，應是大有機會。

On the Hotelling T^2 control chart for the vector autoregressive process

T.-C. Cheng[1], P.-H. Hsieh[2], S.-F. Yang[1]
[1]Department of Statistics, National Chengchi University, Taipei, Taiwan
[2]College of Business, Oregon State University, Corvallis, Oregon, USA
(chengt@nccu.edu.tw, Ping-Hung.Hsieh@bus.oregonstate.edu, yang@nccu.edu.tw)

Abstract - A **vector autoregressive (VAR) model has become a popular multivariate monitoring technique for serially correlated observations often observed in practice. In this article, we examine, via a Monte Carlo approach, the effect of a shift in the model parameter and the sample size in both Phase I and Phase II schemes on control chart statistics, namely, different versions of Hotelling's T^2 when a VAR model is employed. The effects are reported and specific T^2 statistics under various sample sizes is recommended.**

Keywords - **Hotelling's T^2 statistic, vector autoregressive process, statistical quality control**

I. INTRODUCTION

Several models have been proposed to accommodate the interdependence of quality characteristics collected from a multivariate process system, see for example, a state space model by [1] and [2], a Bayesian multivariate local level model by [3], and a multivariate dynamic linear model by [4]. Because of its flexibility in handling serial correlated data, a vector autoregressive (VAR) model has been drawing attention (see, for example, [5] and [6]). By filtering the multivariate process system with a VAR model, one could then monitor the model residuals as a serially independent multivariate series. Subsequently, a typical chart statistics, such as Hotelling's T^2, can be plotted to monitor a certain level of process quality.

Although the use of VAR models is gaining its popularity, an important research question concerning the variation of the resulting T^2 during Phase I of control charting and Phase II of process monitoring requires further investigation. Two major contributing factors are considered in this article: a shift in the model parameter and the sample size. [6] examined theoretically the effect of a change in the model parameter on T^2 and point out the VAR residuals are only asymptotically independently and identically distributed (i.i.d.). Hence, without a sufficiently large sample size, the control limits established in Phase I may have unintended Type I error rate, which in turn, lead to undesirable Type II error rate in Phase II. See [7] for a discussion on the increase of the false alarm rate in a multivariate autocorrelation process. We examine and verify this issue in the simulation study.

II. METHODOLOGY

A. A VAR model

Denote the k variables of a vector autoregressive process $y_t = \{y_{1t}, y_{2t}, ..., y_{kt}\}$ in a k-dimensional vector of response variables. The vector autoregressive (VAR) model with order p is defined as follows:

$$y_t = \mu + (\Phi_1 L + \Phi_2 L^2 + \cdots + \Phi_p L^p) y_t + \varepsilon_t \qquad (1)$$

where L is the lag operator, $\mu_t = \{\mu_1, \mu_2, ..., \mu_k\}$ is the mean vector, and $\varepsilon_t = \{\varepsilon_{1t}, \varepsilon_{2t}, ..., \varepsilon_{kt}\}$ is a k-dimensional vector of the error term. Each Φ_j is an $k \times k$ coefficient matrix for the jth lag, j=1,…,p. ε_t is uncorrelated about time but correlated cross-sectionally. That is, $E(\varepsilon_t \varepsilon_t^{'}) = \Omega$ is invariant about time but may not be a diagonal $k \times k$ matrix. The error term, ε_t, is then assumed as a normal distribution with the mean vector 0 and covariance matrix Ω.

Denote $\hat{\mu}$ and $\hat{\Phi}_j$ as the OLS estimates of μ and Φ_j, respectively, and the estimated systematic model is defined as

$$\hat{y}_t = \hat{\mu} + (\hat{\Phi}_1 L + \hat{\Phi}_2 L^2 + \cdots + \hat{\Phi}_p L^p) y_t, \quad t=1,\ldots,N,$$

where N is the number of observations in the Phase I scheme. The tth residual is then

$$e_t = y_t - \hat{y}_t, \qquad t=1,\ldots,N.$$

The covariance of ε_t is estimated by

$$\hat{\Omega}_1 = \sum_{t=1}^{N} e_t e_t^{'} / N \cdot \qquad (2)$$

If the process is in-control and model (1) is adequate and well estimated, the residuals e_t should also be an asymptotically i.i.d. normal distribution with zero means. The traditional Hotelling T^2 chart can be applied to e_t. For observations y_t in Phase II, [6] showed that the chart statistic is

$$T_t^2 = e_t^{'} \hat{\Omega}_1^{-1} e_t, \qquad (3)$$

that follows a χ^2 distribution with k degrees of freedom.

There are two issues about (2) that require further verification. First of all, $\hat{\Omega}_1$ is a biased estimator for Ω. To obtain an unbiased estimate of Ω, one needs to correct it by a penalty factor on (2) as follows (see [8]):

$$\hat{\Omega}_2 = \sum_{t=1}^{N} e_t e_t^{'} / (N - kp - 1).$$

The use of $\hat{\Omega}_1$ rather $\hat{\Omega}_2$ is due to the fact that Hotelling's T^2 in (3) relies on the likelihood and large sample principles. On the other hand, small sample sizes occur frequently in practice, often distorting the false alarm rate in Phase I, and subsequently, the power in Phase II. Therefore, we consider $k(N-1)/(N-k)F_{k,N-k,\alpha}$, as an alternative critical value to T^2 in (3), where $F_{k,N-k,\alpha}$ denotes an *F* distribution with the degrees of freedoms *k* and *N-k*, and α is the level of significance [9].

B. Statistical quality issues in a VAR process

There are three classes of parameters in a VAR model: the process mean, the covariance matrix of error term, and the autoregressive coefficients of the model. [6] showed that the effect of a parameter shift depends on the magnitude of the shift as well as the resulting estimated parameter values. However, their discussion assumes the control limits are correctly established in Phase I and the uncertainty in parameter estimation due to a small sample size is not accounted for. [5] showed that the false alarm rate may change according to the values of Ω in their in-control VAR(1) simulation study. [6] also pointed out that the VAR residuals are only asymptotically i.i.d. Therefore, to use the convenient Type I error way of setting up control limits for the VAR chart, one needs to have the Phase I sample size large enough in order to guarantee the serial independence of the residuals. However, no conclusion has been drawn with regard to the appropriate sample size. [4] used 100 observations to establish their Phase I control chart, but only the change of **Ω** in the bivariate VAR(1) model was considered. Our simulation will provide more insight into this issue.

As discussed in the previous section, the following Hotelling T^2's and the corresponding control limits are considered in the study:

$$T_{1c}^2 = e_t^{'} \hat{\Omega}_1^{-1} e_t \geq \chi_{k,\alpha}^2 \quad (4)$$

$$T_{2c}^2 = e_t^{'} \hat{\Omega}_2^{-1} e_t \geq \chi_{k,\alpha}^2 \quad (5)$$

$$T_{1F}^2 = e_t^{'} \hat{\Omega}_1^{-1} e_t \geq \frac{k(N-1)}{N-k} F_{k,N-k,\alpha} \quad (6)$$

$$T_{2F}^2 = e_t^{'} \hat{\Omega}_2^{-1} e_t \geq \frac{k(N-1)}{N-k} F_{k,N-k,\alpha} \quad (7)$$

Given α=0.0027, the T^2 control chart schemes yield an overall in-control ARL of approximately 370. Both Phases I and II employ 20,000 replicates to calculate the ARL criterion.

III. RESULTS

A. Simulation design

Any shift on one or more of the three parameters, μ, Φ_j's, and Ω, may change the distributions of the chart statistic and induce out of control signals. To verify the performance of T^2 on monitoring a VAR process, We follow the simulation design in [6]. A tri-variate VAR(3) model is generated with the mean vector $\mu^{'} = (1.1 \quad 2.1 \quad 1.1)$, and the covariance matrix of the error term and the coefficient matrices are defined as follows:

$$\Omega = \begin{bmatrix} 1 & 0.5 & -1.5 \\ 0.5 & 4.25 & 0.25 \\ -1.5 & 0.25 & 2.99 \end{bmatrix}, \Phi_1 = \begin{bmatrix} 0.5 & 0.2 & -0.08 \\ -0.1 & 0.7 & 0.2 \\ 0.25 & -0.22 & 0.54 \end{bmatrix}$$

$$\Phi_2 = \begin{bmatrix} -0.25 & 0.23 & -0.1 \\ -0.34 & 0.32 & 0.12 \\ 0.43 & -0.18 & 0.15 \end{bmatrix}, \Phi_3 = \begin{bmatrix} 0.21 & -0.1 & -0.13 \\ -0.09 & -0.32 & -0.21 \\ 0.15 & 0.05 & 0.48 \end{bmatrix}.$$

The change of mean imposes a multiplier, 1.1, 1.2, or 1.3, of the mean vector above. The shift of covariance matrix uses a multiplier 1.05, 1.1 and 1.2. The change of Φ_j focus on varying Φ_1. That is

$$\Phi_1^* = \begin{bmatrix} 0.5 & 0.2 & -0.08 \\ -0.1 & 0.7 & 0.2 + c \\ 0.25 & -0.22 & 0.54 \end{bmatrix}$$

where *c*=0.2, 0.1, -0.1, -0.2, and -0.3.

Apart from examining the shift in the model parameters, the simulation design considers the effects of various model configurations, specifically, the number of variables or series (*k*), the number of order (*p*), and sample size (*N*). For the Phase I scheme, the sample size, N_1, of each data set is assigned to be 50, 100, 250, and 500. While the sample size for Phase II, N_2, is 20% or 40% of N_1.

B. Simulation results

TABLE I shows the in-control ARL values under the Phase I scheme, indicating the occurrences of false alarms in 20,000 replications. Since the expected ARL is 370, it is obvious that T_{1c}^2 outperforms the other three methods no matter what the sample size, N_1, is. It expects that the sample size increases, the ARL values for all other three methods, T_{2F}^2, T_{1F}^2, and T_{2c}^2 will converge to that of T_{1c}^2 with various converging speeds. The table also shows that the sample size required in Phase I to achieve the theoretical ARL is rather large, and one should be cautious when a more complicated model is employed.

TABLE II presents the out-of-control ARL results under Phase II scheme for a VAR(3) model. "None" in the table denotes the in-control case, which means no

change in any parameter in the Phase II scheme. Under this case, T_{2F}^2 appears supreme even although it does not yield a reasonable performance for small sample size (say N_1 <100) under Phase I scheme. The ARL values show $T_{2F}^2 > T_{2c}^2 > T_{1F}^2 > T_{1c}^2$ in TABLE II.

The change of mean vector results in quite consistent departure of all T^2's in terms of ARL. Under the same N_1 for Phase I and N_2 for Phase II, the comparison of ARL values of these four methods yields $T_{2F}^2 > T_{2c}^2 > T_{1F}^2 > T_{1c}^2$ for different kinds of change of μ in the same model. The ARL values are quite close under the same T^2 no matter what sample size is in the Phase II scheme. For a given T^2, the ARL values increase as the sample size, N_1, in the Phase I increases; however, the ARL values remain similar irrespective of the sample size, N_2, in Phase II.

The smaller out-of-control ARL values should be expected when there exists a shift on any parameter from the process. Nevertheless, we may suspect that the ARL values are too small for some T^2's in TABLE II. In other words, the power of the test statistic is over-stated. To evaluate the accuracy of T^2 in (4)-(7), simulated ARL values for all configurations using large samples are carried out in TABLE III for a VAR(3) model. The values in TABLE III are obtained by using 10000 data points for Phase I and another 10000 observations for Phase II. The same settings for parameters are employed as discussed in the simulation design section. When we look at each row in TABLE III, all the ARL values are quite similar in the same configuration no matter what T^2's are used. It concludes that all versions of T^2 have an equal performance when large samples are applied in both Phase I and II schemes.

To compare the values in TABLE I and TABLE II, T_{2F}^2 is the one fast approaching to the simulated in-control ARL values in both Phase I and Phase II schemes. The T_{1c}^2 suggested by [6] leads to the smallest values among all T^2's in TABLE I and it is very sensitive to the change, which results in an increase in the false alarm rate in Phase II. This phenomenon is also present in the change of mean vector as well as a shift of other parameters and it gets worse as the sample size decreases.

The similar phenomena appear in the change of Ω shown in TABLE II. The larger the multiplier values in the change of Ω, the smaller the ARL values are under the same condition. The ARL values are sensitive to the sample size of N_1 in the shift of variance.

The change of Φ_1 leads to the same results as those of μ and Ω. The ARL values for the change of Φ_1 shows $T_{2F}^2 > T_{2c}^2 > T_{1F}^2 > T_{1c}^2$ in TABLE II. The deviation of ARL values among these T^2's keeps in the right direction as the shift of the parameters, that is, the larger the change is, the smaller the value of T^2 is. Although only one element, *c*, is changed in Φ_1 for a VAR(3) model, the resulting ARL values are very different as shown in TABLE I. The ARL values yielded by T_{2F}^2 are the closest to the simulated out-of-control ARL values (see TABLE III).

IV. DISCUSSION

In summary, all T^2's defined in (4) to (7) are sensitive to the change of parameters and sample sizes. As the sample size in Phase I increases, the ARL values decrease (and converge to the expected ARL value) for all simulation cases under the same conditions. Given the same sample size in Phase I, the in-control ARL values remain quite similar regardless of the sample sizes in Phase II. Although we only consider the change of one parameter at a time, the effect of a parameter shift on ARL is already apparent. One can induce that simultaneous change of the parameters will lead to more significant results.

V. CONCLUSION

We verify, via a Monte Carlo approach, the effects of a parameter shift and sample size on the T^2 statistics when a VAR model is employed. Both the sample size and the critical value for the Hotelling T^2 statistic for Phase I scheme are confirmed to achieve a reasonable average run length, leading to a reasonable Phase II monitoring chart. Our suggestion is to use $\hat{\Omega}_1$ and χ^2 distribution for the Hotelling T^2 statistic under Phase I scheme when sample size is not large, while there is no difference in using (4)-(7) for larger samples. $\hat{\Omega}_2$ and *F* distribution are suggested to be applied under the Phase II scheme.

REFERENCES

[1] A. Negiz and A. Clinar, "Statistical monitoring of multivariable dynamic processes with state-space models," *AIChE Journal*, 43, pp. 2002-2020, 1997.

[2] X. Pan and J. Jarrett, "Applying state space to SPC: monitoring multivariate time series," *Journal of Applied Statistics*, 31, pp. 397-418, 2004.

[3] K. Triantafyllopoulos, "Multivariate control charts based on Bayesian state space models," *Quality and Reliability Engineering International*, 22, pp. 693-707, 2006.

[4] S. I. Chang and K. Zhang, "Statistical process control for variance shift detections of multivariate autocorrelated

processes," *Quality Technology and Quantitative Management*, 4, pp. 413-435, 2007.
[5] A. A. Kalagonda and S. R.Kulkarni, "Multivariate quality control chart for autocorrelated processes," *Journal Of Applied Statistics*, 31, pp. 317-327, 2004.
[6] X. Pan and J. Jarrett, "Using vector autoregressive residuals to monitor multivariate processes in the presence of serial correlation," *International Journal of Production Economics*, 106, pp. 204-216, 2007.
[7] C. M. Mastrangelo and D. R. Forrest, "Multivariate autocorrelated processes: data and shift generation," *Journal of Quality Technology*, 34, pp. 216-220, 2002.
[8] H. Lutkepohl, *New Introduction to Multiple Time Series Analysis*, 2005, Springer, New York.
[9] B. S. Everitt, "A monte carlo Investigation of the robustness of Hotelling's one- and two-sample T^2 tests," *Journal of the American Statistical Association*, 74, pp. 48-51, 1979.

TABLE I
IN-CONTROL ARLs OF PHASE I SCHEME FOR A VAR(3) MODEL

N_1	T_{2c}^2	T_{1c}^2	T_{2F}^2	T_{1F}^2
50	11627.91	880.28	166666.67	5494.51
100	1233.81	531.35	2801.12	1052.63
250	567.67	423.66	730.14	541.54
500	452.90	395.93	510.07	443.71

TABLE II
OUT-OF-CONTROL ARLs OF PHASE I SCHEME FOR A VAR(3) MODEL

Shift		N_2							
		40% N_1				20% N_1			
	N_1	T_{2c}^2	T_{1c}^2	T_{2F}^2	T_{1F}^2	T_{2c}^2	T_{1c}^2	T_{2F}^2	T_{1F}^2
None	50	33.02	15.76	64.99	27.56	33.23	15.91	66.89	27.82
	100	108.30	65.64	173.12	98.29	110.10	66.05	176.29	99.70
	250	232.56	181.97	287.77	223.49	232.29	181.98	287.03	222.62
	500	293.28	257.58	327.33	288.06	291.72	256.11	323.83	286.20
1.1μ	50	22.97	11.73	42.27	19.54	23.05	11.76	42.03	19.61
	100	70.97	44.30	107.12	64.72	70.09	44.02	105.15	63.93
	250	149.11	119.15	182.00	143.93	148.15	118.29	179.86	142.98
	500	189.96	168.46	210.33	186.50	190.51	168.95	211.28	186.93
1.2μ	50	11.85	6.80	19.91	10.31	11.82	6.77	19.89	10.28
	100	31.60	21.21	45.46	29.16	31.58	21.16	45.40	29.10
	250	59.46	49.30	70.42	57.67	59.49	49.41	70.71	57.64
	500	74.83	67.63	81.78	73.70	75.13	67.87	82.18	73.98
1.3μ	50	6.23	4.00	9.48	5.58	6.25	4.02	9.52	5.59
	100	13.30	9.63	17.85	12.52	13.40	9.67	17.92	12.60
	250	23.18	19.73	26.65	22.58	23.28	19.78	26.76	22.69
	500	28.50	26.16	30.71	28.13	28.48	26.18	30.69	28.13
1.05Ω	50	27.94	13.64	53.50	23.49	28.28	13.82	54.36	23.72
	100	86.96	53.69	134.98	79.59	87.72	54.14	137.69	80.06
	250	171.03	136.18	209.03	164.88	172.38	137.63	211.42	166.22
	500	217.60	193.69	241.79	213.94	217.39	193.44	241.84	213.70
1.1Ω	50	24.50	12.31	45.73	20.74	24.41	12.39	45.00	20.68
	100	69.91	44.06	106.02	63.96	70.25	44.16	107.90	64.21
	250	134.07	107.96	162.04	129.16	134.26	108.90	162.28	129.58
	500	165.71	147.35	183.37	162.77	164.95	147.03	182.78	162.11

TABLE II (continued)

1.2Ω	50	18.85	9.89	33.56	16.08	18.94	9.95	33.47	16.13
	100	47.70	31.17	69.41	43.98	48.30	31.55	70.21	44.57
	250	86.03	70.10	102.62	83.29	86.13	69.72	103.08	83.09
	500	101.97	92.03	111.85	100.32	101.97	92.01	111.79	100.35
Φ_1^* -0.2	50	5.65	3.72	8.30	5.09	5.71	3.75	8.40	5.14
	100	11.71	8.66	15.41	11.05	11.68	8.66	15.37	11.01
	250	19.38	16.76	22.04	18.95	19.42	16.80	22.10	19.00
	500	23.12	21.38	24.74	22.84	23.16	21.43	24.78	22.89
-0.1	50	15.15	8.31	26.33	13.05	15.24	8.37	26.50	13.16
	100	42.58	27.69	62.07	39.26	42.25	27.45	61.56	38.94
	250	80.99	66.03	96.58	78.47	81.56	66.28	97.38	78.86
	500	102.32	91.90	112.64	100.60	102.66	92.03	112.76	100.93
0.1	50	11.60	6.81	19.08	10.17	11.53	6.80	18.97	10.09
	100	31.73	21.40	45.06	29.37	31.13	21.09	44.21	28.86
	250	60.50	50.07	71.27	58.63	60.65	50.30	71.55	58.78
	500	74.34	67.02	81.13	73.23	74.25	66.96	81.03	73.18
0.2	50	2.96	2.23	3.90	2.76	2.98	2.24	3.91	2.77
	100	4.87	3.93	5.91	4.67	4.88	3.93	5.91	4.68
	250	7.06	6.33	7.78	6.94	7.06	6.33	7.78	6.94
	500	8.12	7.65	8.54	8.04	8.15	7.68	8.57	8.07
0.3	50	1.39	1.25	1.56	1.35	1.39	1.25	1.56	1.36
	100	1.61	1.48	1.74	1.58	1.61	1.48	1.74	1.58
	250	1.81	1.73	1.88	1.79	1.80	1.73	1.88	1.79
	500	1.89	1.84	1.93	1.88	1.89	1.84	1.93	1.88

TABLE III
SIMULATED ARL VALUES FOR A VAR(3) MODEL

	T_{2c}^2	T_{1c}^2	T_{2F}^2	T_{1F}^2
Phase I	357.143	344.828	357.143	344.828
Phase II				
None	277.778	277.778	277.778	277.778
1.1μ	212.766	212.766	217.391	212.766
1.2μ	84.034	84.034	84.034	84.034
1.3μ	28.490	28.490	28.571	28.490
1.05Ω	285.714	277.778	285.714	285.714
1.1Ω	163.934	163.934	166.667	163.934
1.2Ω	116.279	116.279	119.048	116.279
Φ_1^* -0.2	23.753	23.641	23.810	23.753
0.1	95.238	94.340	96.154	95.238
0.1	76.923	76.336	76.923	76.923
0.2	8.097	8.078	8.117	8.097
0.3	1.868	1.867	1.870	1.868

國科會補助計畫衍生研發成果推廣資料表

日期:2011/11/29

國科會補助計畫	計畫名稱: 2×2列聯表之穩健診斷
	計畫主持人: 鄭宗記
	計畫編號: 99-2118-M-004-002- 學門領域: 其他應用統計

無研發成果推廣資料

99 年度專題研究計畫研究成果彙整表

計畫主持人：鄭宗記	計畫編號：99-2118-M-004-002-
計畫名稱：2×2 列聯表之穩健診斷	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	1	1	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 ■未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

We propose two kinds of robust estimators to provide resistant results for analyzing the logistic regression model. The proposed approaches are able to deal with binary responses and categorical covariates. One of these estimators can also be used to identify outlying cells for several 2x2 contingency tables, which contribute some insights to the related strand in the literature. The real data examples implementing the proposed approaches yield some new findings.

The current study can be extended to the K 2xJ tables for the analysis of case-control studies with the exposure measured at several levels, and considers model misspecification for the proportional odds regression model.